The ASCI Blue Mountain System

3TOps Capability at Los Alamos National Laboratory

This leaflet describes the Blue Mountain System at Los Alamos National Laboratory (LANL). The system is a component of the Accelerated Strategic Computing Initiative (ASCI) program, a collaboration between DOE Defense Programs and the Sandia, Lawrence Livermore, and Los Alamos National Laboratories. The goal of the ASCI program is to create and utilize leading-edge capabilities in simulation and computations modeling. In an era without nuclear testing, these computational goals are vital for maintaining the safety, reliability, and performance of the nation's nuclear stockpile.

To meet the needs of stockpile stewardship in the year 2010, modeling and simulation applications must achieve validated higher-resolution, three-dimensional, complete-physics, and full-system capabilities. This level of computation requires high-performance computing (HPC) far beyond our current level of performance. LANL's Blue Mountain System is one step in furthering these computational goals.

Overview and Performance

The system consists of 48 Silicon Graphics Origin 2000 shared memory multi-processor computers with 128 250-MHz processors on each machine (total of 6144 processors). The system was installed over a period of a few months as the Origins were delivered in several phases. ASCI Blue was available to users throughout these phases of its configuration buildup, providing an excellent test environment for the scalability of applications. Full configuration was reached in November of 1998.

The cluster has a composite of 1.5 Terabytes of RAM and 76 Terabytes of fiber channel disk. The system represents a peak capacity of 3.072 Teraflops (3 trillion floating point mathematical operations) per second, making Blue Mountain one of the most powerful computer systems installed on site in the world.

Performance achievements on Blue Mountain have been impressive. Soon after the full system was installed, the industry standard LINPACK benchmark ran on 5040 of the processors with a rate of 1.608 TeraOps. Within a month of installation, most of LANL's ASCI applications were successful in scaling to over 6000 processors. These applications routinely run on 2 to 3 thousand processors.

The ASCI Blue Mountain System

Networking for ASCI Blue

Perhaps the most critical issue for computing on ASCI Blue is the networking required to connect the individual machines into an integrated parallel compute engine. Using HIPPI 800 interconnects, LANL has pushed parallelism into the network. Each SMP currently has a dozen 800 Mbit bi-directional HIPPI channels, providing throughput of over a gigabyte per second when shipping data between SMPs. Over the entire system, achievable aggregate throughput is over 50 gigabytes per second.

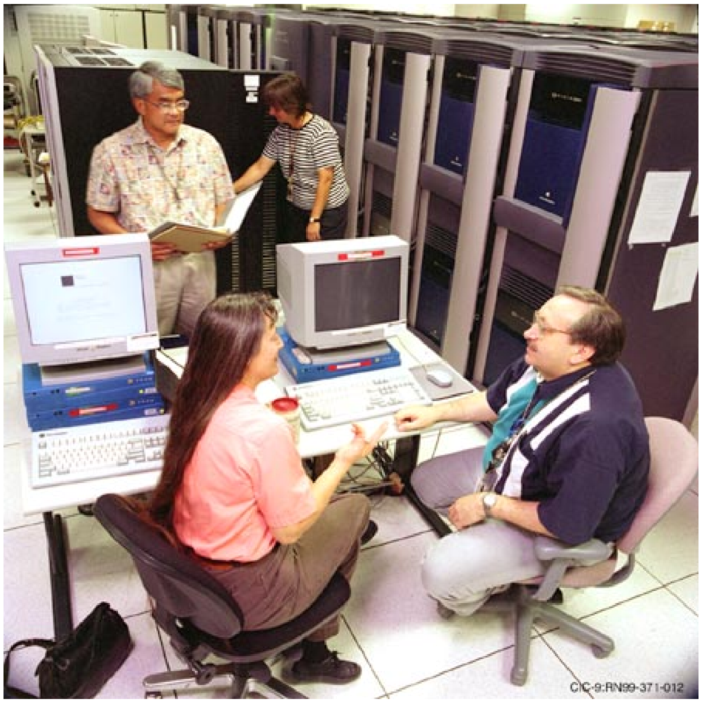
LANL is working with SGI and other vendors to push standards-based networking even higher. The Gigabyte System Network (GSN), to be released from SGI in early 2000, will vastly increase the communication capability of the cluster. A single GSN link will carry eight times as much information as a HIPPI link. Each SMP will have six GSN links, all operating in parallel. In addition to improved bandwidth, GSN will provide lower latency over the network using the ST protocol. The result will be an even more capable system, with balanced performance between computational speed and networking capability.

Software

For the high performance needed by ASCI applications, the multiple SMPs must be used together as a single system. This is primarily accomplished via the Message Passing Interface (MPI) software. MPI uses the HIPPI 800 OS bypass, a low-level protocol that achieves low latency. The objective is to write portable applications using MPI but to optimize performance through the use of OS bypass. To further optimize performance, LANL is also writing a library that will use OS bypass without MPI. In performance comparisons, the MPI library has provided a bandwidth of 90 MB/second sustained with 144 microseconds one-way latency and the OS bypass library has given 140 MB/second bandwidth with 104 microseconds one-way latency.

The Load Sharing Facility (LSF) software from Platform Computing Corporation is used for job scheduling and control on the system. LSF distributes jobs across the 48 machines of ASCI Blue using features such as queue or machine limits, queue priorities, processor reservation, and job backfilling to provide efficient utilization of the system. In addition to queueing batch jobs, the software allows interactive work spanning multiple machines of the system, a capability that facilitates the development and testing of applications. LSF allows quick responses to changing programmatic needs through relatively simple configuration modifications, often requiring no changes in users' job submittals.

Archival storage for the ASCI Blue Mountain system is provided by the High Performance Storage System (HPSS), which is a new generation storage system for extremely large amounts of data (petabytes) with the ability to access data at very high data rates (10s to 100s Mbytes/sec). ASCI applications on Blue Mountain have generated HPSS files ranging up to 340 GB. Today the storage on HPSS is approaching 100 Terabytes, 8 times that of a year ago. The current growth rate is 13TB/month, 4 times the growth rate of a year ago.



Support

A team of approximately 45 people, involving both LANL employees and SGI personnel, supports the Blue Mountain 3TOps system. The work of this team involves extensive systems integration, tying together system management, networking, security, distributed resource management, data storage, applications support, development of parallel tools, user consultation, documentation, problem tracking, usage monitoring, operations, and facilities management.

Future Milestones for ASCI

The successful integration of the Blue Mountain 3TOps system represents one milestone on the road to scaling applications and supporting a fully operational simulation capability for stockpile stewardship. Building upon the experience and knowledge gained with the 3Tops system, LANL will procure and install a computational system that will achieve a peak performance level of 30 TeraOps by mid-year 2001. It is expected that a 100 TeraOp capability is needed by the year 2004 in order to meet the goals of stockpile stewardship.

For more information about ASCI Blue, contact:

Don McCoy (dmccoy@lanl.gov or 505-667-0940),

John Morrison (jfm@lanl.gov or 505-667-6164), or

Manuel Vigil (mbv@lanl.gov or 505-667-5243).

Work performed under the auspices of the U.S. DOE by LANL under contract W-7405-ENG-36